TrueWay: A Highly Scalable Multi-Plane Multi-Stage Buffered Packet Switch

H. Jonathan Chao, Jinsoo Park, Sertac Artan, Shi Jiang
Department of Electrical and Computer Engineering
Polytechnic University, Brooklyn, NY 11201
chao@poly.edu, jspark118@yahoo.com, {sartan01, sjiang01}@utopia.poly.edu

Guansong Zhang
Fortinet, Inc.
Sunnyvale, CA 94085
gzhang@fortinet.com

1. Introduction

To keep pace with Internet traffic growth, researchers have been continually exploring new switch architectures with new electronic and optical device technologies to design a packet switch that is cost-effective and scalable to a very large capacity, e.g., a few hundred Tera bps or even a few Peta bps [1]. A new packet switch family, called Birkhoff-von-Neumann two stage load balancing switch, has been proposed and received much attention [2][3]. This family of switches do not require a centralized packet scheduler and thus is highly scalable. The challenging part is to maintain packets' sequences. Another alternative to scale the switch is to use the multi-plane multi-stage buffered architecture. For instance, Cisco's CRS-1 system [4] based on Benes network can scale up to 92 Tbps.

We have also independently prototyped a similar ultra-scalable multi-plane multi-stage buffered switch architecture, called TrueWay, based on the Clos network [5]. We have investigated various packet scheduling schemes, as well as link-to-link, and port-to-port flow control schemes to improve the system performance. Several challenging design issues related to designing the TrueWay switch are listed below.

- How to efficiently allocate and share the limited on-chip memories?
- How to intelligently schedule packets on multiple paths while maximizing the memory utilization and system performance?
- How to minimize link congestion and prevent buffer overflow (i.e., stage-to-stage flow control)?
- How to maintain packets' orders if they are delivered over multiple paths (i.e., port-to-port flow control)?

This paper addresses the above issues and provides several solutions for each issue. We have shown by simulation that the TrueWay switch with a speedup of 1.6 is able to perform nearly as well as the output buffered switch under most interested traffic distributions. A small-scale switch fabric prototype has been built on a 16-card chassis with high-speed SerDes interconnections at the backplane (with 640 Gbps capacity), and with FPGA chips on each card to reconfigure the switch to test various packet scheduling schemes. With today's ASIC technology, using for instance, 64x64 switch chip with SerDes interfaces and VCSEL (Vertical Cavity Surface Emitting Laser) optical interconnections, the TrueWay switch can scale up to 40Tbps.

2. Switch Architecture

Fig. 1 shows the TrueWay switch architecture. The ingress traffic manager (TMI) distributes packets to different paths of different switch planes while the egress traffic manager (TME) collects packets traversing the switch fabric and buffers them to be transmitted to the network. The switch fabric consists of *p* switch planes, each a three-stage Clos network.

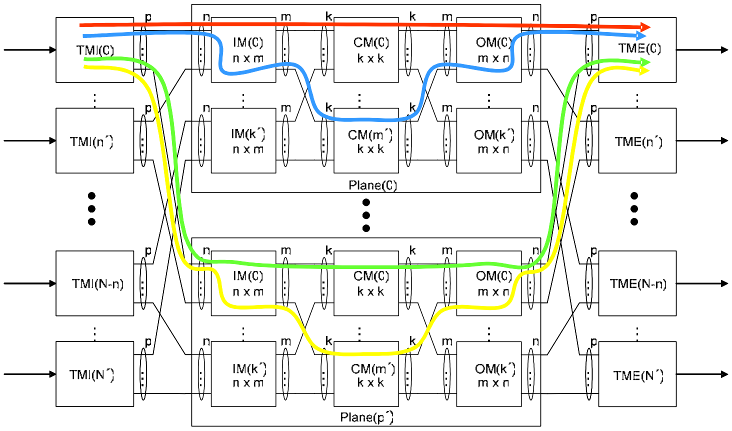

Figure 1. TrueWay Switch Architecture

The modules in the first, second, and third stages are denoted as input modules (IMs), center modules (CMs), and output modules (OMs). Each module can be logically considered a cross-point buffered switch. The first stage of the switch plane consists of *k* IMs, each with *n* inputs and *m* outputs. The second stage consists of *m* CMs, each with *k* inputs and *k* outputs. The third stage consists of *k* OMs, each with *m* inputs and *n* outputs.

Similar to conventional packet switches, each incoming packet is segmented into multiple fixed-length cells. The cells are then routed through the switch fabric independently and reassembled into packets using reassembly queues at the output ports. Cells are served using strict priority: Only when cells in the higher priority queues are completely served, can the cells in the lower priority queues be served.

3. Packet Scheduling and Flow Controls

A. Packet Scheduling

In a multi-plane multi-stage switch, there are more than one path between any input-output pair, thus cells that belong to the same flow may easily get out-of-sequence. Typically, the number of reassembly queues at each TME is equal to the product of the number of inputs, the number of paths between any input output pair, and the number of priorities. To reduce the number of reassembly queues, and to offer high throughput and fairness, while maintaining cell integrity, we investigated four packet scheduling schemes for the TrueWay switch. They are (1) cell interleaving (CI), (2) complete packet interleaving (CPI), (3) partial packet interleaving (PPI), and (4) dynamic packet interleaving (DPI). We analyzed and evaluated each scheme thoroughly in terms of throughput, fairness and hardware complexity.

B. Link-to-Link Flow Control

A flow control mechanism is implemented to make sure there is no cell loss in the switch fabric. It is a common practice to have a large memory at line cards and a small memory at switch fabric connection points. When a cell is transmitted between the switching stages, the receiver should have free memory space to store the cell until it is transmitted to the next stage. Since the receiver has limited memory space, if its memory becomes full, the sender must hold the cells until the receiver has free space. We studied two well known flow control schemes: Back-Pressure and Credit-Based Flow Control, also known as N23 [6]. And we propose a new flow control scheme dedicated to the TrueWay switch, called DQ scheme.

C. Port-to-Port Flow Control

In a multi-plane multi-path switch, cells may travel through different paths and experience different queuing delays. Thus, packets can be delivered out-of-sequence through the switch fabric. To maintain packet order in the TrueWay switch, we studied four port-to-port flow control schemes. They can be categorized into two approaches namely hashing method and buffer resequencing method. In the first approach, hashing, we force the cells belonging to the same flow to take the same path through the switch fabric. As a result, all cells from a given flow will experience the same amount of queuing delay. Thus packets are delivered in order. Along this approach, we present two hashing schemes: static hashing and dynamic hashing. On the other hand, the buffer resequencing approach allows packet out-of-sequence within the switch fabric. However, TME uses a resequencing buffer to resequence the cells back in order before delivering them to the next link. We studied two practical resequencing techniques for the TrueWay switch. They are time-stamp-based resequencing and window-based resequencing.

4. Testbed

We prototyped a small-scale TrueWay switch on a chassis, as shown in Fig. 2, to evaluate various packet scheduling schemes, as well as link-to-link and port-to-port flow control schemes. The prototype consists of a high-speed multi-trace backplane and up to 16 plug-in cards. The backplane and plug-in cards were manufactured using a state-of-the-art 14-layer PCB process. The backplane provides a total bandwidth of 640 Gbps for the inter-connection of the plug-in cards, where each card has 16 2.5-Gbps duplex ports connecting to other cards. Each of the plug-in cards can have up to two units, where each unit consists of one Xilinx Virtex-II 3000 FPGA, one Velio Octal 3.125 Gbps SerDes, and the necessary glue logic.

The chassis provides a generic framework for testing different packet scheduling schemes under the Clos-network switch structure by programming FPGA chips. The hardware platform can be configured, for instance, to a 16×16 switch with 4 switch planes and a total capacity of 40 Gbps.

The current prototype only consists of a 4×4 switch with 2 switch planes for demonstration purposes. This switch consists of five cards and six FPGAs in total. The traffic manager (TM) card has two FPGAs, where each FPGA can accommodate two TMI/TME unit pairs (i.e., two TM chips), making a total of four TMs connecting to the four ports of the switch. The switch fabric includes four cards, each with one FPGA. One card/chip can accommodate two switching modules (SMs). Each IM/OM card has one IM/OM pair (i.e., one SM chip) and each CM card has two CMs (i.e., one SM chip). One IM/OM card/chip and one CM card/chip form a single switch plane. Fig. 3 shows a photo of this implementation. We also implemented testing features into the TM chips. Each TM chip has a built-in traffic generator that can generate different types of packet streams to be applied to the switch. At the same time, TM chips also check if any packet is lost in the switch fabric.

The Xilinx ISE Foundation toolkit with VHDL was used for the entire FPGA implementation. A 2-unit card is shown in Fig. 3.

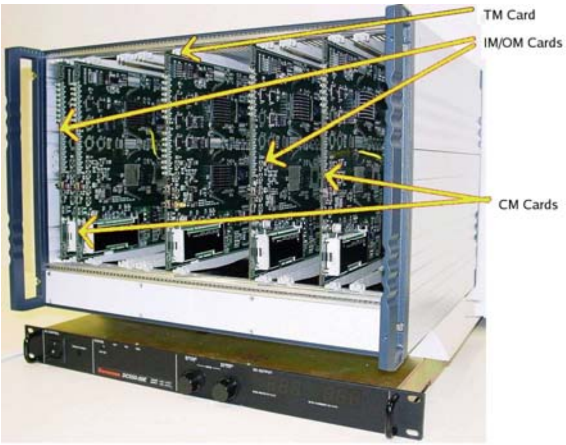

Figure 2. TrueWay Testbed

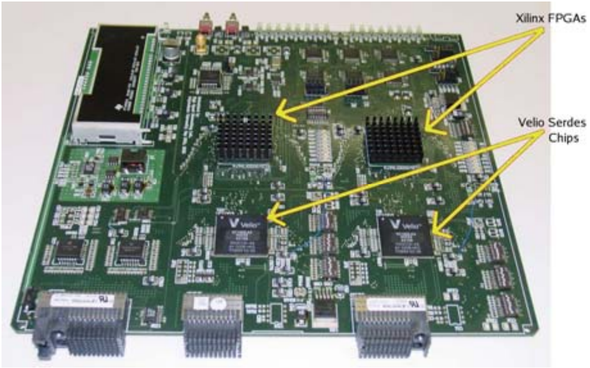

Figure 3. Testbed card with 2 FPGAs (under heat sink) and 2 Velio Serdes+

5. Future Work

A. *Multicast Capability*

Multicast function is traditionally implemented with a multicast bitmap in the cell header (i.e., at cell level) or a multicast table in the switch fabric (i.e., at flow level). But neither of them scale to a large switch, such as the TrueWay switch.

The first approach requires a large bitmap for a large-scale switch. For example, when building a 40 Tbps system, the required bitmap size is 128bit (64bit for CM and 64bit for OM), which causes too much overhead in cell headers. The second approach requires a multicast table which is too big to fit into the on-chip memory when the switch size is large. Since each OM can receive a packet from any TMI through any CM in the same plane, the number of flows for an OM can be more than half million. In order to support this large number of multicast flows, each OM requires a memory size larger than 16Mbits, which is challenging to include in the switch module chip.

Therefore, finding a feasible and efficient multicast scheme for the TrueWay switch is a topic of future research.

B. *Fault Tolerance*

When a fault occurs in the switch system (due to the failure of components or interconnection wires cut), the switch system should be able to detect, isolate, and restore the fault as soon as possible without any service disruption. By taking advantage of multiple plans and multiple paths, the TrueWay switch shall be able to reconfigure the system gracefully. The techniques for detection/isolation/reconfiguration require for future study.

Reference:

[1] H. J. Chao, “Next Generation Routers,” Invited Paper, *Proceedings of the IEEE*, Vol. 90, No. 9, Sep. 2002.

[2] C.S. Chang, D.S. Lee, and Y.S Jou, “Load Balanced Birkhoff-von Neumann Switches, Part I: One-stage Buffering,” *Computer Communications*, vol. 25 pp. 611-622, April 2002.

[3] I. Keslassy, S.T. Chuang, K. Yu, D. Miller, M. Horowitz, O. Solgaard, and N. McKeown, “Scaling Internet Routers Using Optics,” in *ACM SIGCOMM* Aug. 2003, Karlsruhe, Germany.

[4] Cisco Systems, http://www.cisco.com/en/US/products/ps5763

[5] C. Clos, “A Study of Non-Blocking Switching Networks,” *Bell Sys. Tech. Jour.*, pp. 406-424, March 1953.

[6] H. T. Kung, Trevor Blackwell, and Alan Chapman, “Credit-Based Flow Control for ATM Networks: Credit Update Protocol, Adaptive Credit Allocation, and Statistical Multiplexing,” in Proc. of *ACM SIGCOMM*, pp. 101-115, 1994.